DOCUMENT RESUME

ED 400 754 HE 029 594

AUTHOR Greenwald, Anthony G.
TITLE Applying Social Psychology to Reveal a Major (But Correctable) Flaw in Student Evaluations of Teaching.
PUB DATE Mar 96
NOTE 34p.; Paper presented at the Annual Meeting of the American Psychological Association (103rd, New York, NY, August 11-15, 1995).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Course Evaluation; Course Selection (Students); *Data Interpretation; Grade Inflation; *Grading; Higher Education; Reliability; Statistical Analysis; Student Attitudes; *Student Evaluation of Teacher Performance; Teacher Effectiveness; *Teacher Evaluation; Teacher Promotion; Validity
IDENTIFIERS *University of Washington

ABSTRACT

Higher education relies on student ratings to evaluate faculty teaching, partly because the alternatives (expert peer appraisals or objective performance criteria) are costly or unavailable. Because student ratings are crucial not only to improving instruction, but also in making or breaking faculty careers, it is important to assure that they provide valid indications of instructional quality. Analyses of large data sets obtained at University of Washington show that student ratings are prone to artifacts that can produce occasional substantial underestimates of teaching ability for instructors who grade strictly (and overestimates for those who grade leniently). Some likely system impacts of this distortion of ratings are to nudge (1) instructors toward lenient grading, and (2) students toward nonchallenging courses. The bright side of this picture is that the usefulness of student ratings can be improved statistically. While it has been found that giving inflated grades produces inflated ratings and higher student workloads generally produce lower ratings, statistical adjustment of data, removing invalid variance, can derive more accurate ratings. The appendix contains the Instructional Assessment System form used for faculty evaluation at the University of Washington. (Contains 27 references.) (Author/JLS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

APPLYING SOCIAL PSYCHOLOGY TO REVEAL A MAJOR (BUT CORRECTABLE) FLAW IN STUDENT EVALUATIONS OF TEACHING

ED 400 754

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Anthony G. Greenwald
University of Washington

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Anthony Greenwald
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

Abstract. Higher education relies on student ratings to evaluate faculty teaching, partly because the alternatives (expert peer appraisals or objective performance criteria) are costly or unavailable. Because student ratings are crucial not only to improving instruction, but also in making or breaking faculty careers, it is important to assure that they provide valid indications of instructional quality. Analyses of large data sets obtained at University of Washington show that student ratings are prone to artifacts that can produce occasional substantial underestimates of teaching ability for instructors who grade strictly (and overestimates for those who grade leniently). Some likely system impacts of this distortion of ratings are to nudge (a) instructors toward lenient grading, and (b) students toward nonchallenging courses. The bright side of this picture is that the usefulness of student ratings can be improved statistically.

Acknowledgment. Material in this article was presented as an address for the Donald T. Campbell Award from the Society of Personality and Social Psychology, presented at the 1995 meetings of the American Psychological Association in New York. The research summarized here was greatly facilitated by the Office of Educational Assessment at University of Washington, and was conducted in collaboration with Gerald M. Gillmore, who will be co-author of more detailed reports. Additional support was provided by Grant SBR-9422242 from National Science Foundation, and Grant MH 41328 from National Institute of Mental Health. For comments on various drafts and material preliminary to this one, the author thanks Robert D. Abbott, Philip C. Abrami, Kenneth A. Feldman, Gerald M. Gillmore, Joe Horn, George S. Howard, Herbert W. Marsh, Scott E. Maxwell, Jeremy D. Mayer, Robert S. Owen, Lloyd K. Stires, and John E. Stone. Correspondence may be addressed to the author at Department of Psychology — Box 351525, University of Washington, Seattle, WA 98195-1525, and electronic mail to agg@u.washington.edu.

HE 029 594

BEST COPY AVAILABLE

APPLYING SOCIAL PSYCHOLOGY TO REVEAL A MAJOR (BUT CORRECTABLE) FLAW IN STUDENT EVALUATIONS OF TEACHING

I remember *very* clearly when my research interest in teaching evaluations began. In the first week of April, 1990, I received the summary of student ratings from my Winter quarter honors undergraduate seminar in social psychology. I had taught this same seminar in 1989, the previous year, and had received the highest average ratings that I had ever received at University of Washington. In 1989, my ratings for this course were in the top 10% of University of Washington faculty. Having taught the course in 1990 according to the same plan as the previous year, I naturally expected similarly high ratings. Imagine my surprise when my 1990 ratings turned out instead to be the lowest ratings that I had yet received at University of Washington, placing me in the 2nd lowest decile of the university's faculty.

Did I immediately think, "Wow! Those 80 or 90% above me must be really great teachers!"? No.

Did I think, "Wow! Here's a great opportunity for some research!"? No.

Did I think, "Wow. What on earth did I do wrong?" Yes.

I was not only surprised by the ratings, but quite upset. I could take some comfort from having received much higher ratings a year earlier for what I believed to be exactly the same course. Even so, getting ratings near the bottom of the distribution for faculty at my university was painful. How much more painful is it when a new junior faculty member receives similarly low ratings, perhaps for the very first course taught, and knowing that those ratings are likely to be considered in an eventual tenure decision? Having spoken with junior faculty members who were in exactly that position, I know that, beyond being upset and disappointed, they will begin to search with some urgency for things they can do to improve their ratings. But what should they change?

A Thought Experiment

Imagine that you are that junior faculty member. You have just taught a course for the first time and have received low ratings. In addition to your disappointment with the low ratings, you were also disappointed with students' performances on examinations. They did poorly on questions based on material that, you thought, had been well covered in your lectures. You are about to teach the same course again, and are convinced that you need to change something. Consider two options. One option is to blame yourself, deciding that you did not explain the material clearly enough; you can correct that by spending more time on basic material, trying to assure that students will master at least that basic material. The other option is, in effect, to blame the students, deciding that they didn't work hard enough; you can oblige them to work harder by giving weekly paper assignments or quizzes.

Both strategies are likely to raise students' grades. The retreat-to-basics approach will increase grades at least partly by reducing coverage of course material, so that less work will be needed to achieve whatever percentage level of mastery is required for a given grade. The more-frequent-evaluation approach will not make it easier to earn a given grade, but should nevertheless get students to achieve more by prodding them to spend more time on the course.

Both approaches may also lead to improved ratings (see Powell, 1977). Nevertheless, the retreat-to-basics alternative may be much more likely to be adopted, as a consequence of two likely sources of influence. One influence is from written comments that typically accompany low ratings; these are likely to include complaints that tests covered material that was never clearly explained. The second likely influence is from advice provided by colleagues who draw on their own and others' experience with student ratings. In order to get an idea of the richness of advice about student ratings that is available from colleagues, consider the following recent internet bulletin-board message.

> Students who think they are getting As tend to think more highly of their professor than students who believe they are getting Cs. So for a professor to maximize evaluations, the best bet is to give out a softball midterm, so that everyone thinks they're getting a great grade. However, if a professor really wants students to learn, the ideal method is to give a hard midterm, and scare the students into studying. Thus, the goals of pedagogy and high instructor evaluation are in direct opposition. If you give out lots of Cs and students think you are a great professor, you're probably excellent. If you give out all A and A minuses, and students think you're just OK, you probably suck.[1]

Are Ratings Contaminated by Grades?

Any consideration of strategies to increase ratings is likely to quickly focus on the very simplest strategy that is suggested by academic folklore — just give higher grades. The strategy of giving high grades is so very tempting if only because it is so very simple. One need make no change beyond recalibrating the course's grade scale. To judge from anecdotes available on the academic grapevine, the faith that the grade-increasing strategy works appears to have some basis in real experience of teachers. Nevertheless, it would be very desirable to have a methodologically sound research answer to the question, *If I give higher grades, will I get higher ratings?*

Historical Trends in Research on Student Ratings

An electronic search for publications on student ratings reveals that the possible effect of grades on ratings was the subject of much research, peaking 15 to 20 years ago. Figure 1 characterizes a sample of that research in the period from 1971 to 1995. It can be seen that, over the entire 25-year period, more publications favored validity than invalidity. However, the research changed sharply in character around 1980.

---

[1] Abridged and quoted from an electronic mail message circulated by Jeremy D. Mayer, Department of Government, Georgetown University, July 11, 1995.

Figure 1. Shifting appraisals of validity of student ratings. This plot summarizes the author's categorization of study conclusions, inferred from their abstracts retrieved from electronic searches of PsycINFO and ERIC, using for both data bases the search query, *(student rating$1 or teaching evaluation$1) and (bias or valid$3 or invalid$3)*. The *$n* suffix includes in the search any words found by appending up to *n* letters after the stem. 'Biased' indicates study conclusions that student ratings of instruction are contaminated by some source of invalidity. The ERIC search was limited to unpublished reports, in order not to have the two searches produce duplicates.

As can be seen in Figure 1, 1980 marked the beginning of a decline in research activity on student ratings. However, the analysis of conclusions shows that this was a decline specifically of studies that remained neutral (dropping from 31 to 16 between 1976-1980 and 1981-1985) and those that were critical (dropping even more drastically, from 15 to 3). At the same time, the number of studies supporting validity remained the same, and these increased in proportion from a minority of 35% (25/71) to a majority of 57% (25/44). By the 1990s, research on validity of ratings had diminished to such a low level that it is easy to infer that earlier contributions had resolved the major issues. Articles published from about 1980 on do indeed give the impression that some major questions about ratings validity were considered to have been answered. Researchers were willing to describe the validity of student ratings in rather general ways, including assertions that grades were unlikely to produce bothersome influences on ratings.

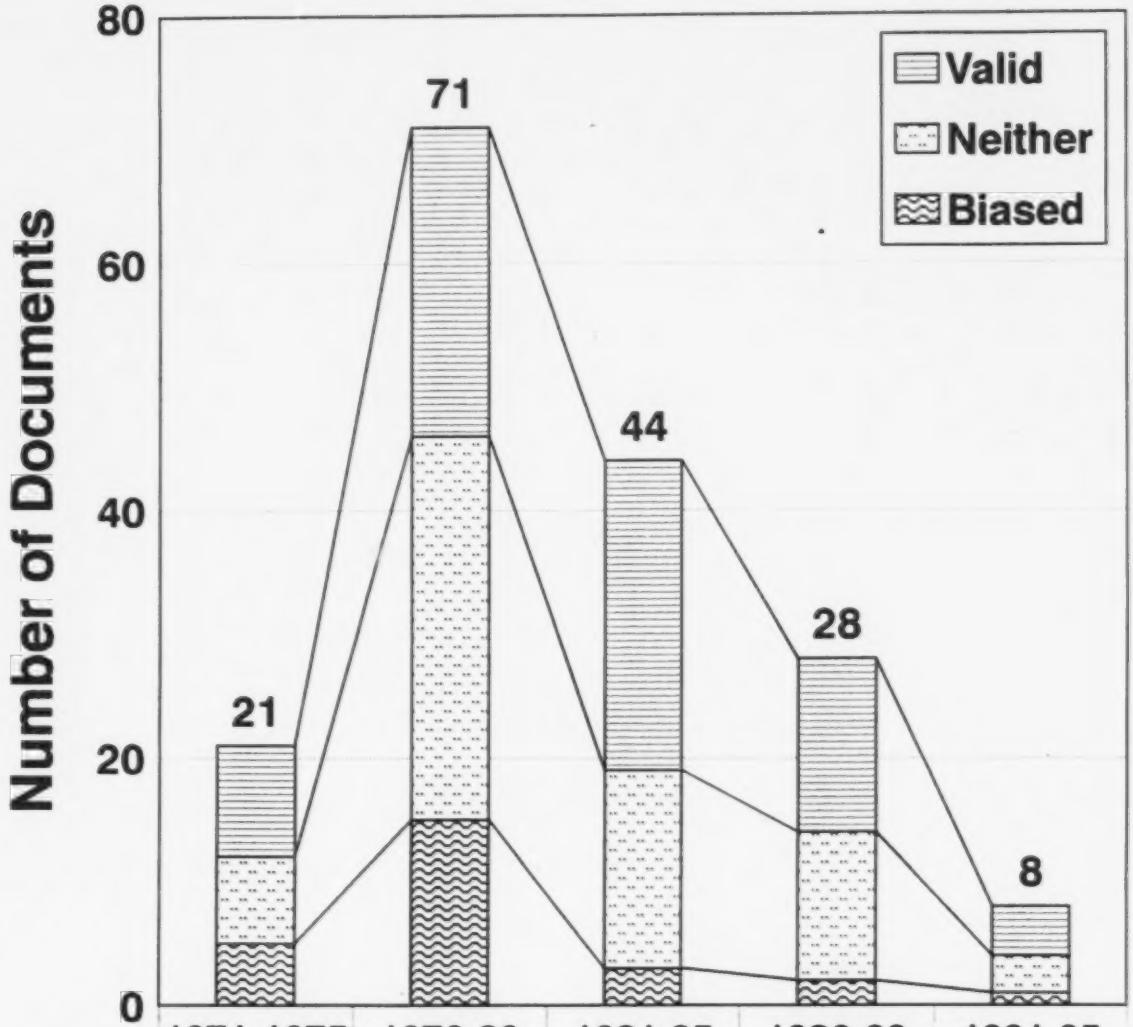
> In general, . . . most of the factors [that] might be expected to invalidate ratings have relatively small effects. . . . Some studies have found a tendency for teachers giving higher grades to get higher ratings. However, one might argue that in courses in which students learn more the grades should be higher and the ratings should be higher so that a correlation between average grades and ratings is not necessarily a sign of invalidity. . . . My own conclusion is that one need not worry much about grading standards within the range of normal variability. (McKeachie, 1979, pp. 390, 391)

> Probably, students' evaluations of teaching effectiveness are the most thoroughly studied of all forms of personnel evaluation, and one of the best in terms of being supported by empirical research. ... [A]lthough it is possible that a grading leniency effect may produce some bias in student ratings, support for this suggestion is weak and the size of such an effect is likely to be insubstantial in the actual use of student ratings. (Marsh, 1984, pp. 749, 741)

> [Recent] evidence has suggested ... that rather than signaling possible contamination and invalidity of student evaluations, the observed relation between grades and student ratings might reflect expected, educationally appropriate relations. (Howard, Conway, & Maxwell, 1985, p. 187)

> In general, student ratings tend to be statistically reliable, valid, and relatively free from bias or the need for control; probably more so than any other data used for evaluation. (Cashin, 1995, p. 6).

These quotes acknowledge that grades and ratings are correlated, but also express the judgment that this correlation can and should be interpreted without concluding that grades create a bothersome contamination of ratings.



	1971-1975	1976-80	1981-85	1986-90	1991-95
Valid	9	25	25	14	4
Neither	7	31	16	12	3
Biased	5	15	3	2	1

Five-Year Periods

1970s: Research Questioning Validity of Student Ratings

Although research published in the 1970s covered a variety of concerns about validity, a major concern of that period was the possible effect of grades on ratings. The concern with grade-induced bias is apparent in the following quotes.

> The present evidence, then, supports a notion that a teacher can get a "good" rating simply by assigning "good" grades. The effect of obtained grades may bias the students' evaluation of the instructor and therefore challenges the validity of the ratings used on many college and university campuses. (Snyder & Clair, 1976, p. 81)

> The implications of the findings reported are considerable, and it is suggested that the validity of student evaluations of instruction must be questioned seriously. It is clear that . . . an instructor [who] inflates grades . . . will be much more likely to receive positive evaluations. (Worthington & Wong, 1979, p. 775).

These were conclusions from experiments in which grades had been manipulated upward or downward, and the manipulated grades were observed to raise or lower student ratings, correspondingly. There were several such experiments, mostly appearing in the 1970s, that had been conducted in actual undergraduate courses (Chacko, 1983; Holmes, 1972; Powell, 1977; Vasta & Sarmiento, 1979; Worthington & Wong, 1979). On reading these field experiments side by side in the 1990s, it is easy to conclude that, in combination, they make a rather powerful case that ratings can be biased sharply by arbitrary grading practices. Those experiments are difficult to repeat in the 1990s, because their grade manipulations imposed stresses and used deceptions that university human subjects review committees do not now look kindly upon. However, the best argument for not replicating these experiments 20 years later is that it hardly seems necessary to do so — the results of the older studies were clear enough so that there seems little doubt about what new replications would find.[2]

So, this is a strange situation. On the one hand, experimental results reported during the 1970s appeared to demonstrate that grading practices influence student ratings. Contemporary folklore among academicians also endorses the conclusion that one can raise ratings by inflating grades. On the other hand, concern about the possibility that grading practices can distort student ratings largely disappeared from the scholarly literature on student ratings after about 1980. How did research manage to quiet concerns that ratings could be biased by manipulating grades?

---

[2]Contemporary reviews of student ratings literature either omit treatment of these natural classroom experiments on effects of manipulated grades on ratings, or mention them only in the context of suggesting that they are collectively flawed (e.g., Marsh & Dunkin, 1992, p. 202).

BEST COPY AVAILABLE

1980s: Demonstrations of Convergent Validity of Student Ratings

Since 1980, research on student ratings has mostly been in the form of correlational construct validity designs. Three kinds of studies provided evidence that has supported the construct validity of student ratings.

Multisection validity studies. In the best of the largest group of construct validity studies, multiple sections of the same course are taught by different instructors, with student ability approximately matched across sections and with all sections having identical or at least similarly difficult examinations. Using examination performance as the criterion measure of achievement, these studies have determined whether differences in achievement for students taught by different instructors are reflected in the student ratings of the instructors. The collection of multisection validity studies has been reviewed in several meta-analyses. Although the meta-analytic reviews do not agree on all points concerning the validity of student ratings, nevertheless it is clear that multisection validity studies yield evidence for modest validity of ratings. Correlations between ratings and exam-measured achievement average about 0.40 (see the overview of meta-analyses by Abrami, Cohen, & d'Apollonia, 1988, esp. pp. 160-162).

Multisection validity studies favor construct validity of ratings by supporting an interpretation of observed grades-ratings correlations in terms of common effects of a third variable, teaching effectiveness. If grades correlate with ratings simply because good teachers produce both high grades and high ratings, then all is well with the validity of student ratings.[3]

Path-analytic studies. The second type of correlational construct validity study also explores the idea that effects of third variables on both grades and ratings explains their correlation, but considers third variables other than teaching effectiveness. For example, Howard and Maxwell (1980) applied path analysis techniques to show that grades and ratings were both related to measures of students' level of motivation for courses, from which they concluded that

> the relationship between grades and student satisfaction might be viewed as a welcome result of important causal relationships among other variables rather than simply as evidence of contamination due to grading leniency. (p. 810)

In another example of this type of study, Marsh (1980) observed that

> A path analysis demonstrated that students' Prior Subject Interest had the strongest impact on student ratings [and] accounted for about one-third of the relationship between Expected Grades and student ratings. . . . Expected Grade was seen as a likely bias — albeit a small one — to the ratings, and even this interpretation was open to alternative interpretations. (pp. 219, 236)

---

[3] Interpretation of this approximate .40 correlation as reflecting processes other than, or in addition to, validity of student ratings has also been suggested. For example, Marsh and Dunkin (1992, pp. 173ff.) note that this correlation could be contributed to either by motivational variations among students in different sections or by greater student satsifaction with higher grades.

BEST COPY AVAILABLE

Multitrait-multimethod studies. The third type of construct validity study seeks to demonstrate that student ratings possess both convergent and discriminant validity — that is, to demonstrate that they correlate (a) relatively well with measures based on other methods for assessing the construct of quality of instruction, and (b) relatively less well with measures assumed to assess other constructs (e.g., Freedman, Stumpf, & Aguanno, 1979; Howard, Conway, & Maxwell, 1985; Marsh, 1982). Such multitrait-multimethod studies typically have reported evidence for both convergent and discriminant validity of student ratings, although they usually have done so without considering expected grades as a source of contamination.

Overview: The Question of Discriminant Validity Remains

There is an Emperors' Clothes quality to the research literature on validity of student ratings. The researchers of the 1970s, who demonstrated experimentally that grade manipulations affected ratings, declared in effect that the student ratings emperor had a wardrobe problem. Researchers of the mid 1970s to mid 1980s, who reported construct validity studies, concluded that the emperor was in fact clothed. If one reads carefully the latter construct validity studies, it becomes apparent that they did not declare the emperor to be *fully* clothed. The question of what was left exposed translates, in construct validity terms, to the question of discriminant validity of student ratings. Construct validity studies have established that student ratings do, to a moderate extent, measure what they're supposed to measure. But we want to know also how well they avoid bias resulting from sensitivity to things that they're not supposed to measure — which is to say that we want to know about their discriminant validity.

When there is good discriminant validity, having only modest convergent validity means that one has an unbiased, even if noisy, measure. For example, think of weighing people on a scale that will produce a value somewhere within 10 pounds of their correct weight. If a series of these weights has an independent normal distribution that is centered on the correct weight, then one can get a very good measure simply by being patient enough to take multiple readings and average them. If student ratings have moderate convergent validity accompanied by good discriminant validity, one might be reluctant to treat individual-course ratings as highly accurate, especially ratings obtained from small classes, but one should not otherwise be concerned.

The situation is importantly different when moderate convergent validity is accompanied by some failure of discriminant validity. Consider, for example, what happens when economists report seasonally adjusted monthly indexes of unemployment. The raw figure of percent of people out of work fluctuates in response to seasonal factors such as the influence of weather on building construction schedules and farm harvests. These systematic fluctuations are irrelevant to the overall state of the economy, and make the raw unemployment rate misleading and somewhat invalid as an indicator of economic health. Fortunately, this discriminant validity problem of the raw unemployment rate does not render it useless. If one applies a correction for the time of year, then the adjusted unemployment rate provides a considerably more valid measure of the overall economy.

9 BEST COPY AVAILABLE

Student ratings measures are now used in most undergraduate institutions without any adjustments. In other words, student ratings are being treated as if they have excellent discriminant validity, meaning that they have no substantial contaminating influences. On the one hand, this seems implausible, because convergent validity with nonratings measures of quality of instruction has never been shown to be more than moderate, and also because replicated experiments, conducted in actual instructional settings, have demonstrated that grading policy variations substantially affect student ratings. On the other hand, however, for the past 15 years well respected researchers have asserted that it is acceptable to treat student ratings as construct-valid measures of instructional quality. This is a paradox.

Findings and Theories

An acceptable response to such a paradox is to attempt to subdue it with theoretical analysis and new data. Toward that goal, a series of data collections was conducted at University of Washington between 1992 and 1994. These data were obtained using a new rating form (Gillmore & Greenwald, 1994) that had been developed partly to facilitate research that could improve the usefulness of student ratings. Data were obtained from multiple samples, each of a few hundreds of courses that were diverse in subject matter, size, and academic level, but were also self-selected by virtue of instructors having volunteered to use the new rating form.

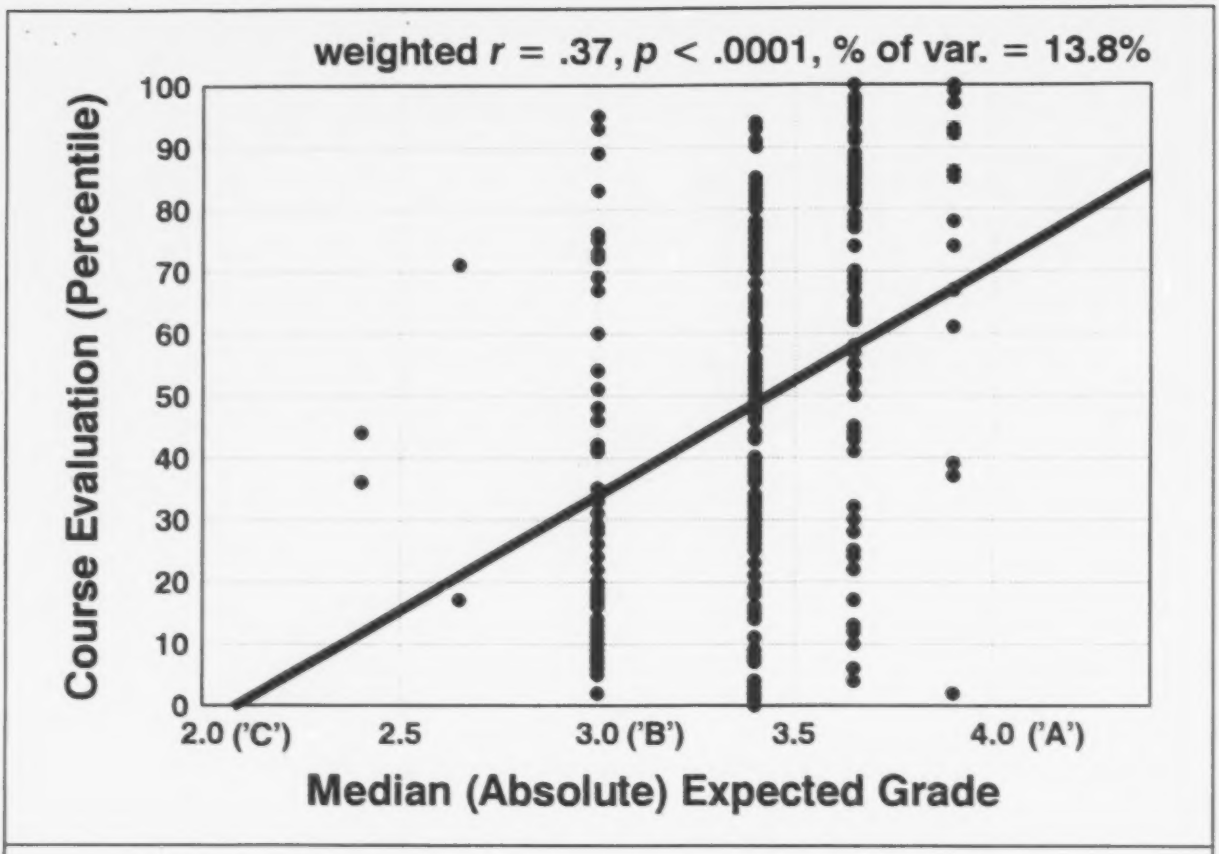
Five Grade-Related Data Patterns in Student Ratings

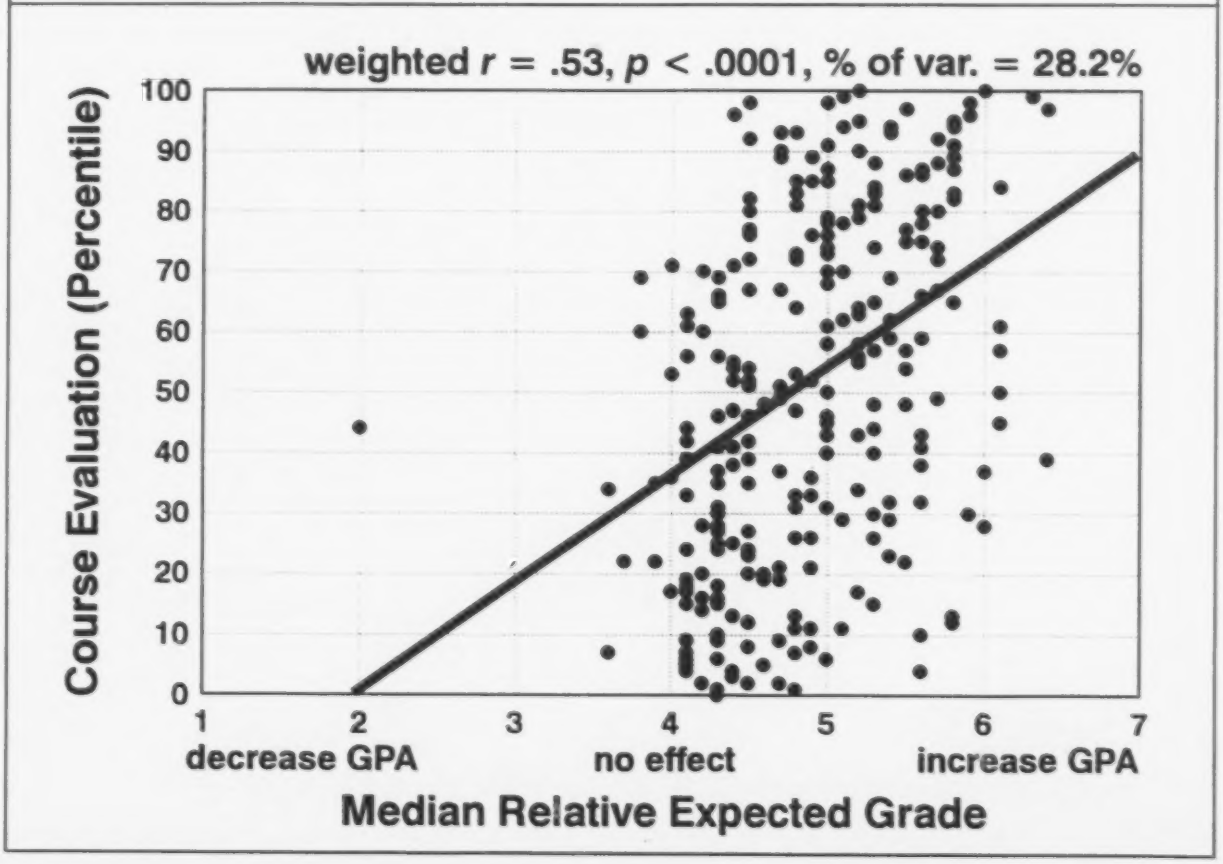
With the exception of one finding that was tested only during a single academic term (the fourth one listed below), the following five findings have been corroborated in separate data collections over three or more academic terms in university-wide samples of courses at University of Washington. The first two findings are ones that were also previously obtained in numerous other studies. The remaining three are the more novel contributions of the University of Washington studies.

1. **Positive grades-ratings relationships between classes**. Figure 2 shows the relation between grades and ratings from a sample of courses at University of Washington. The ratings measure is an average of two subscales based on a total of 18 rating items. Eleven of these items described characteristics of the instructor and were averaged into an overall "instructor" subscale. The other seven items described aspects of the student's achievements in the course and were averaged into a "self/progress" subscale. (The form containing these items is provided as the Appendix.)

The first panel of Figure 2 shows the regression function that related the average of these two ratings subscales to median expected grade of a sample of undergraduate courses. This expected grade measure is similar to ones that appear frequently in studies of student ratings. The second panel of Figure 2 plots a similar relationship, but uses a different and novel measure of expected grades. For the second measure of expected grades, students were asked to compare their expected grade to the average of their grades in other courses. This novel measure assesses the expected effect of the course grade on the student's grade-point average, and it is therefore

BEST COPY AVAILABLE





BEST COPY AVAILABLE

referred to as a *relative expected grade* measure. (By contrast, the more familiar measure that was used for the first panel of Figure 2 can be called *absolute expected grade*.) The between-class relationship between grades and ratings shown in Figure 2 is quite a reliable observation that has previously been reported in many studies (see Stumpf & Freedman, 1979, for an overview). Figure 3 shows this grade-ratings relationship in the form of a structural equation model that relates the two ratings subscales to the two expected grade measures.

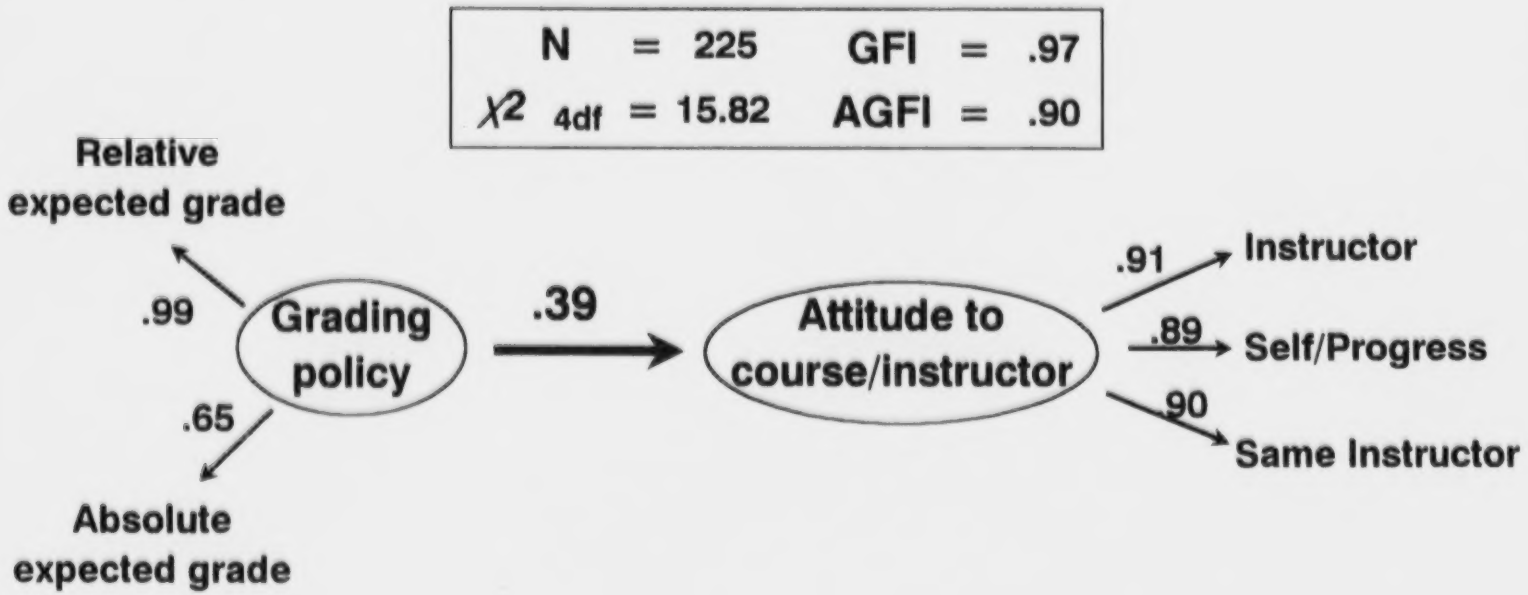
Figure 2. .Regressions of overall course attitude on two measures of expected grades. These are between-course analyses, meaning that courses are the units of analysis. These data were obtained from a university-wide sample of 254 courses in the Winter term of 1994 at University of Washington, and the analyses used number of respondents in each section as a weight. The criterion measure of evaluation, and the absolute (upper panel) and relative (lower panel) expected grade predictors are described in the text. Regression slopes are superimposed on the scatterplots.

Figure 3. Structural model including two measures of instructor grading policy and three measures of attitude toward the course and instructor. The 'Instructor' measure is averaged from Items 1-11, the 'Self/Progress' measure is averaged from Items 12-18, and the 'Same Instructor' measure is Item 23 of the University of Washington Form X (see Appendix). The positive between-course relationship between Grading Policy and Attitude to Course/Instructor is measured by the +.39 value of the path linking those two latent variables. Analysis was limited to 225 courses from Winter 1994 that had at least 10 respondents. GFI = goodness of fit index; AGFI = goodness of fit adjusted for degrees of freedom.

2. **Positive grades-ratings relationships within classes**. Grades-ratings correlations such as those shown in Figures 2 and 3 are routinely also obtained *within* classes (see Stumpf & Freedman, 1979, for an overview). In the University of Washington data, the between-class relationship is characterized by a larger regression slope than the within-class grades-rating relationship (the ratio of regression slopes is approximately 1.9:1). At the same time, the within-class relationship accounts for about twice as much variance in ratings (approximately 16%) as did the between-class relationship (7%) when the two predictors were used simultaneously to predict individual students' ratings in a large, multicourse data set. The within-class grades-ratings relationship is important because it cannot be explained as the effect of any variations in instructor's teaching ability — the instructor is a constant within any class.

3. **Stronger grades-ratings relationships with relative (than absolute) measures of expected grade**. Figure 2 showed that the grades-ratings relationship was stronger when students were asked the question about expected grade in the relative-grade form, which invoked comparison to their performance in other classes. In regression analyses that predicted ratings simultaneously from both the absolute and relative expected grade measures, it was found repeatedly that the relative grade question yielded a very substantial gain in percent of ratings variance explained, over and above that explained by the absolute expected grade question. By contrast, the absolute grade measure accounted for virtually nothing beyond what was explained by the relative grade measure. These patterns were apparent in both between-course and within-course analyses (see Greenwald & Gillmore, in preparation). The comparison of relative and absolute grade measures was a novel feature of the University of Washington research. Consequently, the finding that the grades-ratings correlation is stronger for the relative-grade measure is a new one.

BEST COPY AVAILABLE


N = 225
GFI = .97
χ2 4df = 15.82
AGFI = .90
Relative expected grade
.99
Grading policy
.65
Absolute expected grade
.39
Attitude to course/instructor
.91
Instructor
.89
Self/Progress
.90
Same Instructor

4. **Grade-related halo effect in judging course characteristics**. In Winter Quarter of 1994, approximately 100 instructors at University of Washington agreed to add a small set of items to their regular rating forms. The added items included three that requested judgments of course characteristics that were consdered unlikely to have more than a small relation to the quality of instruction in a course. These three items sought students' judgments of (a) legibility of instructor's writing, (b) audibility of instructor's voice, and (c) quality of classroom facilities to aid instruction (such as an overhead projector). Figure 4 shows magnitudes of grades-ratings correlations for these three items both between and within courses. There was no evidence of a grades-ratings relationship in the between-courses analysis, consistent with the assumption that these items are peripheral to instructional quality. However, the within-courses analysis showed clear positive relationships. Although these within-courses relationships were smaller than within-courses relationships observed for the instructor and self/progress scales, they were nevertheless extremely stable statistically. When it is considered that all students in the same classroom saw the same instructor's handwriting, heard the same instructor's voice, and had the same classroom teaching aids, the observation of these within-sections relationships is remarkable. The content of items on which these grade-halo effects occurred — especially their noncentrality to most conceptions of instructional quality — suggests the potency of grade influences on student ratings.[4]

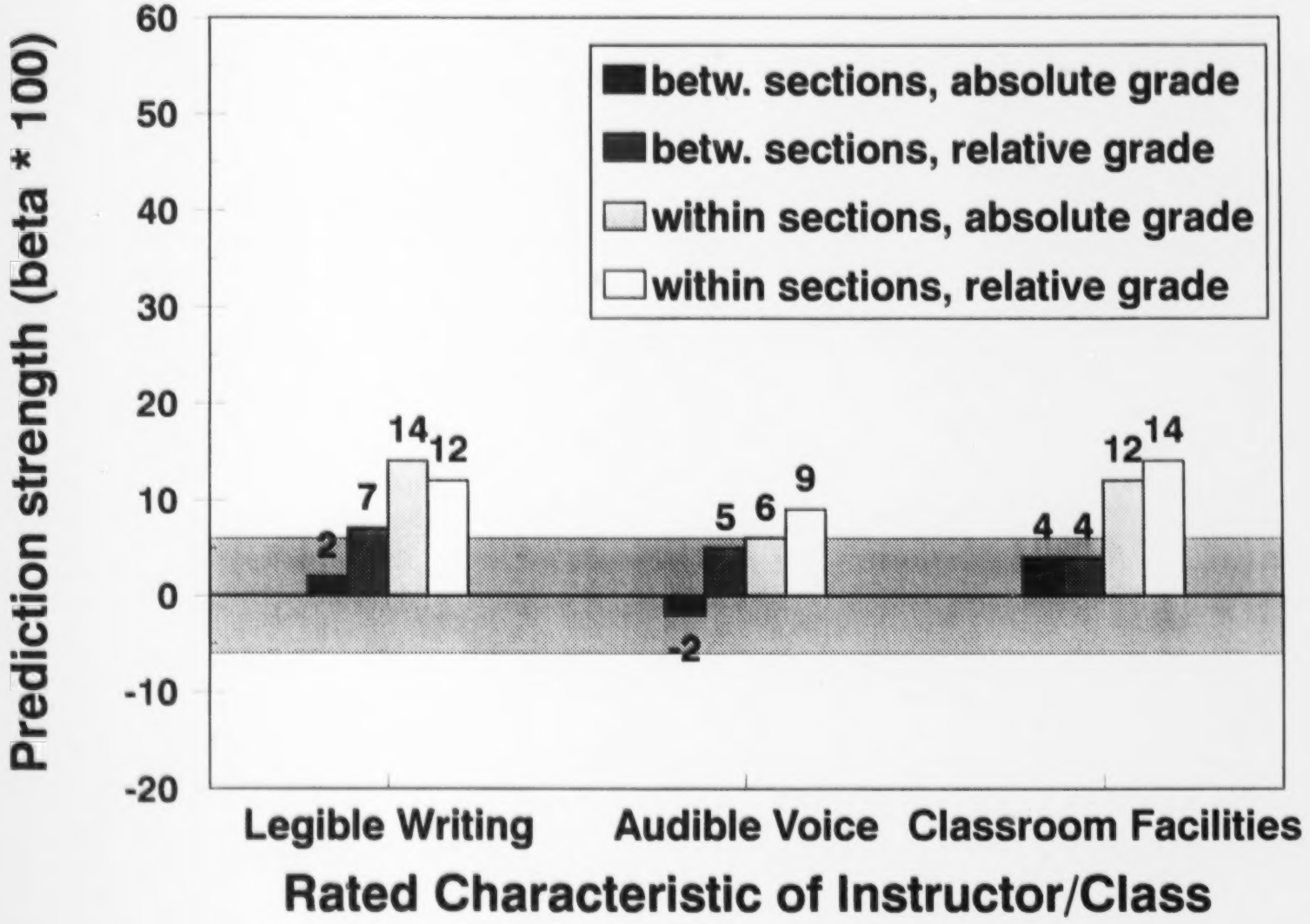
Figure 4. Effect of grades on items that appear peripheral to the construct of quality of instruction. Results are reported as beta coefficients, which provide an effect size measure. Data are from 66 courses (those that had data from more than 10 respondents) at University of Washington in Winter 1994. Total Ns ranged from 1588 to 1610 for the various analyses. The shaded region includes beta values that should not be considered different from zero by a conservative statistical criterion ($\alpha = .005$, 2-tailed).

5. **Negative grades-workload relationship between classes**. It seems reasonable to expect that students should work harder in courses in which they receive high grades than in ones in which they receive low grades. The reasonableness of this expectation rests on two assumptions: (a) that grades awarded in a course provide an indicator of student achievement or learning in the course, and (b) that students work harder in courses in which they learn much than in courses in which they learn little. These two assumptions lead directly to the expectation that students should work harder in courses that give high grades than in courses that give low grades. However, in data obtained repeatedly at University of Washington, this expected positive relationship between grades and course workload was not found. To the contrary, the data repeatedly revealed a substantial negative relationship between course grades and workload — students reported doing more work in courses that had low expected grades than in courses that had high expected grades. This relationship, based on data obtained in the Winter term of 1994 (and found equally clearly in other terms) is shown in the structural equation model of Figure 5. Although tests of the expected grades-workload relationship have not frequently been reported in previous research,

---

[4]Previous findings that front-of-class seating is associated with higher grades (e.g., Knowles, 1982) provide the basis for a possible student-motivation interpretation of the within-courses relationships of expected grades to ratings of instructor voice and legibility, although not the relationship to ratings of classroom facilities. I thank Lloyd K. Stires (personal communication, October 26, 1995) for noting the relevance of the classroom seating variable to these data.

BEST COPY AVAILABLE


Prediction strength (beta * 100)
60
50
40
30
20
10
0
-10
-20
betw. sections, absolute grade
betw. sections, relative grade
within sections, absolute grade
within sections, relative grade
2
7
14
12
-2
5
6
9
4
4
12
14
Legible Writing
Audible Voice
Classroom Facilities
Rated Characteristic of Instructor/Class

other studies have indeed observed the same surprising negative relationship between expected grades and workload in between-course analyses (e.g., Marsh, 1980, pp. 234-235).

Figure 5. Structural equation model using the same data set modeled in Figure 3. The 'Challenge,' 'Good Hours,' and 'Total Hours per Credit' measures are based, respectively, on Items 20, 26, and 27 of Form X (see Appendix). The negative between-course relationship between Grading Policy and Workload is measured by the -.32 value of the path linking their latent variables. GFI = goodness of fit index; AGFI = goodness of fit adjusted for degrees of freedom.

Five Theories of the Grades-Ratings Correlation

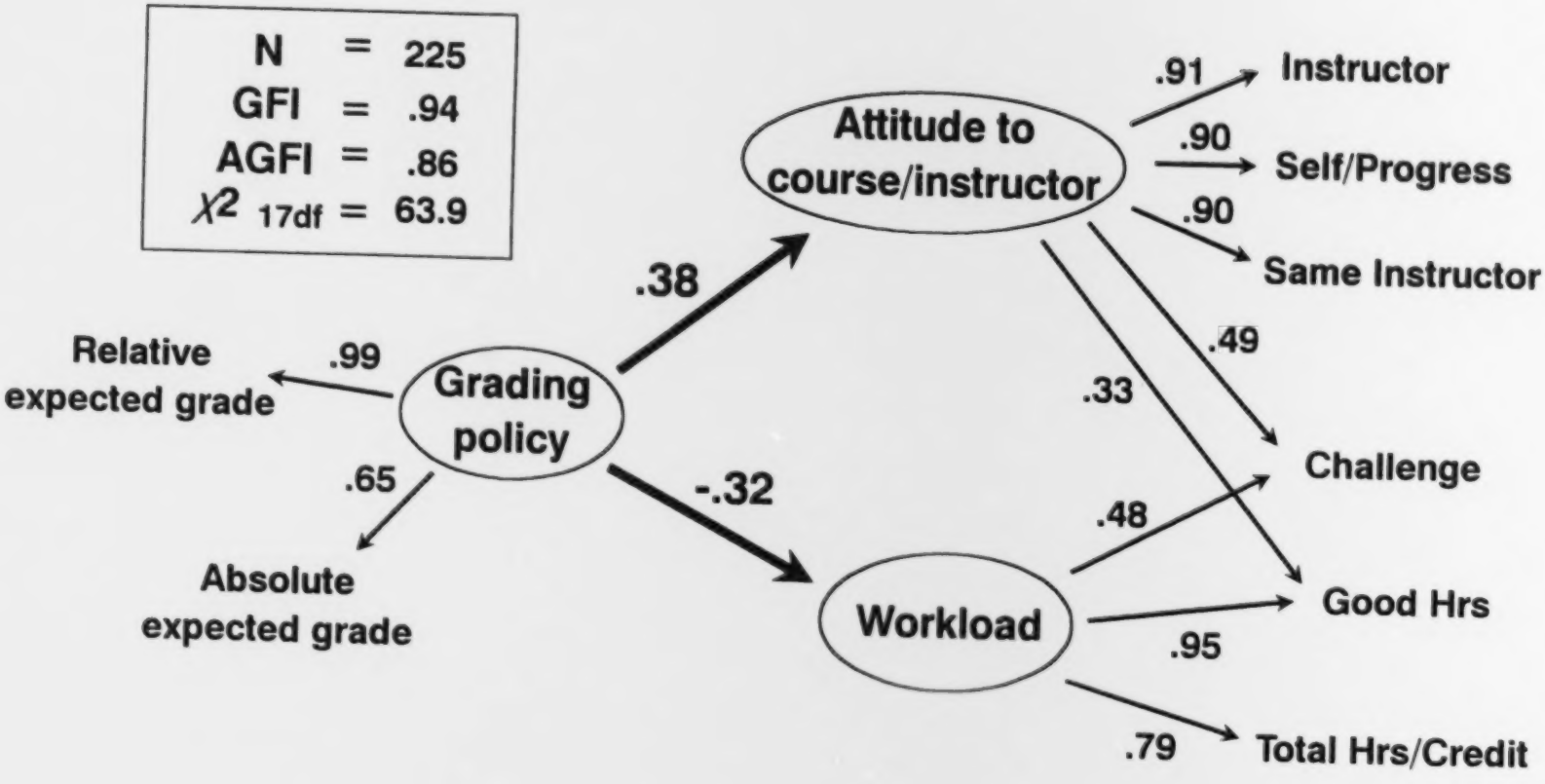
The following five theories vary in the level of construct validity that they credit to student ratings as measures of instructional quality. The first three theories explain the positive grades-ratings correlation by assuming that a third variable influences both grades and ratings. By appealing to third variables, these three theories avoid the assumption of a causal influence of grades on ratings, which is the discriminant-validity-undermining theme of the remaining two theories.

1. **Teaching effectiveness influences both grades and ratings**. This is the one theory that is fully based on the presumed construct validity of student ratings. The central principle of the teaching effectiveness theory is that strong instructors teach courses in which students both (a) learn much (therefore they earn and deserve high grades), and (b) give appropriately high ratings to the course and instructor. In the teaching effectiveness theory, instructional quality is thus the third variable that gives the positive grades-ratings correlation an interpretation fully consistent with construct validity of student ratings measures of instruction.

2. **Students' general academic motivation influences both grades and ratings**. In this theory, the correlation between grades and ratings is credited to variations in students' motivation. Compared to unmotivated students, students with strong academic motivation should both (a) do better in their course work and (b) more fully appreciate the efforts of the instructor, possibly even inspiring the instructor to superior performance. Because students within any course should vary in their level of academic motivation, this theory can explain the grades-ratings correlation within courses. It can also explain the between-course grades-ratings correlation by assuming that courses differ in their success in attracting highly motivated students. For example, courses that have a reputation of being difficult are likely to be taken only by highly motivated students. This student motivation theory has received frequent favorable mention in the research literature on student ratings (e.g., Howard & Maxwell, 1980; Marsh, 1984).

3. **Students' course-specific motivation influences both grades and ratings**. This theory supposes that any student's motivation is variable on a course-by-course basis rather than being a fixed characteristic of that student. This variant of the motivation theory is useful because it can explain the increased strength of the grades-ratings correlation when expected grades are assessed in the relative-grade form (see Figure 2). Both of the motivation theories imply less than perfect construct validity of student ratings because they credit the relation between grades and ratings to a characteristic of students, rather than to a characteristic of instructors. However, to the extent

BEST COPY AVAILABLE


N = 225
GFI = .94
AGFI = .86
X2 17df = 63.9
Attitude to course/instructor
Grading policy
Workload
.91
Instructor
.90
Self/Progress
.90
Same Instructor
.49
Challenge
.48
Good Hrs
.95
.79
Total Hrs/Credit
.33
.38
-.32
.99
Relative expected grade
.65
Absolute expected grade

that student motivation can itself be credited to characteristics of the instructor, construct validity is retained by these motivation theories.

4. **Students give high ratings in appreciation for high grades**. The central idea of this theory is that praise induces liking for the praiser (especially if the praise is greater than expected — see Aronson & Linder, 1965). The translation of this familiar social psychological principle into the ratings context is that the instructor in effect praises the student via a high grade, and the student's return liking is expressed by giving the course and instructor a high rating. This theory has been the focus of much controversy in past research on validity of student ratings, where it is usually identified with the labels of *leniency* or *grade satisfaction*. The leniency interpretation was strongly advocated by researchers who were critical of ratings validity in the 1970s, but its support appeared to diminish greatly in the wake of the correlational construct validity research conducted in the late 1970s and early 1980s. Most mentions of leniency or grade-satisfaction theories in post-1980 publications are in the context of asserting that leniency may account for only minor and ignorable influences on student ratings (see previous quotations of such conclusions). At the same time, the leniency theory appears to have achieved some credibility in academic folklore. This underground support can be seen in the speed with which an instructor who has received low ratings is likely to receive informal advice that raising students' grades (or perhaps just raising their expectations about grades) can help to solve the problem.

5. **Students infer course quality and own ability from received grades**. Social psychological *attribution* theories hold that people make inferences both about their own traits and about the properties of situations in which they act by observing the outcomes of their actions. Research in the attribution theory tradition has established that a favorable outcome for one's own behavior typically leads to the inference that one possesses desirable traits, whereas an unfavorable outcome is more likely to induce perceptions of situational obstacles to success. A simple summary of these attributional principles is that people tend to accept credit for desired outcomes while denying responsibility for undesired outcomes (Greenwald, 1980). Applying these principles to the academic context yields the expectation that high grades will be self-attributed to intelligence and/or diligence, and low grades to poor instruction. Social psychological attribution theory matured after the peak period of research activity on student ratings, which perhaps explains why this type of interpretation has seen less discussion than some others in research on student ratings. Some recent discussion of attribution interpretations of student ratings can be found in papers by Gigliotti and Buchtel (1990) and Theall, Franklin, and Ludlow (1990); see also the recent overview by Feldman (in press).

Evaluation of the Five Theories

The relative success of the five theories in dealing with the set of five findings is summarized in Table 1, and discussed here by reconsidering the five findings.

Table 1. Success of five theories in explaining five patterns in student-ratings data.

BEST COPY AVAILABLE

Table 1

Success of Five Theories in Explaining Five Patterns in Student-Ratings Data

Type of explanation	Hypothesis	Positive between-class grades-ratings correlation	Positive within-class grades-ratings correlation	Relative > absolute grade effect	Grade effect radiates to peripheral items (halo)[a]	Negative between-class grades-workload correlation
Third variable affects both grades and ratings	Third variable is instructor's teaching effectiveness	✔	✘	✘	✘	✘
	Third variable is student's general academic motivation	✔	✔	✘	✘	✘
	Third variable is student's course-specific motivation	✔	✔	✔	✘	✘
Grades influence ratings	Leniency: Students reward/punish instructors who give high/low grades	✔	✔	✔	✔	na
	Attribution: Grades provide information about course quality and student ability	✔	✔	✔	✔	na

Notes: ✔ = hypothesis predicts result; ✘ = hypothesis predicts either a null or opposite-direction result; na = hypothesis does not bear on the result.

[a]This halo effect is a positive grade-ratings correlation (across students, within courses) for items that, rationally, should be evaluated in the same way by all students in the same class (i.e., independently of their grades).

Between-class grades-ratings correlation. Of course, all five theories explain the between-class grades-ratings correlation, which was a necessary requirement in order for each to have earned membership in the set of theories under consideration. The teaching effectiveness theory shows up as weakest of the five theories, because it accounts for nothing beyond the between-class grades-ratings correlation.

Within-class grades-ratings correlation. Because, in the teaching effectiveness theory, the variable that influences both grades and ratings is a constant (the instructor) within any classroom, that theory cannot explain the covariation of grades and ratings within any class. By contrast, the two third-variable theories that allow student differences within a classroom to be related to ratings are able to explain the within-class grades-ratings correlation. Also, the two grades-influence-ratings theories very directly explain why students who get high grades provide the highest course ratings.

Greater grades-ratings correlation for relative-grade measure. The teaching effectiveness interpretation does not explain any within-class grades-ratings correlation, let alone the greater strength of this correlation for the relative-grade than the absolute-grade measure. The finding that ratings are best correlated with the extent to which expected performance deviates from the student's general level of performance also creates difficulties for the general academic motivation theory's expectation that ratings should be associated with the student's assumed stable level of motivation. By contrast, the course-specific motivation theory and the two grades-influence-ratings theories are able to explain why ratings associated with a specific grade are higher when that grade is a relatively high one for the student than when it is the student's typical grade.

Radiating halo effect. All three of the third-variable theories should expect data patterns at odds with the halo effects shown in Figure 4. For the teacher effectiveness theory, if there are any grade effects on the legibility, audibility, and class facilities items, those effects should appear in between-class analyses (but they don't) and they should not appear in within-class analyses (but they do). The two student-motivation third-variable theories are strained in attempts to account for the pattern of grade-related effects on these three items. To spell this out: One might suppose that highly motivated students are more likely to read the instructor's handwriting easily, to hear the instructor clearly, and perhaps even to notice the classroom facilities. Given either student-motivation interpretation, however, these effects should have appeared in between-courses analyses, as well as within courses. The two social psychological theories that credit grade influences on ratings to irrational motivated judgment processes are quite consistent with radiation of the halo effect to peripheral judgments.

Negative correlation between grades and workload. As mentioned previously, the negative grades-workload relationship indicates a flaw in at least one of two assumptions on which the expectation of a positive relationship rested. The first assumption was that the expected grade in a course provides a satisfactory measure of student learning from the course. The second was that students learn more from courses that demand more work. Because rejecting the second assumption so blatantly defies common sense, it seems likely that the first assumption is in error. Rejecting the first assumption — that expected grades are satisfactory indicators of student learning — is damaging to all three of the third-variable theories, which share the assumption that grades

and ratings both reflect the construct of student achievement. The two social psychological grades-influence-ratings theories also do not account for the negative correlation between grades and workload. However, that shortcoming does not embarrass these two theories, because they have no stake in assuming that grades and ratings converge on the construct of student achievement.

Explaining the Negative Grades-Workload Relationship: Tough versus Tender Teachers?

The negative between-course correlation between grades and workloads is the one finding for which no satisfactory (or at least plausible) theoretical explanation has yet been suggested. The following is a speculative attempt, based on the assumption that instructors vary along a dimension that might be labeled leniency-strictness. This dimension is conceived by assuming that relatively lenient instructors teach easy courses and give high grades, whereas relatively strict instructors teach difficult courses and give low grades.[5] When, as is typically true at colleges and universities, students are free to choose many of their courses, such variations in leniency can produce undesirable consequences. In particular, if students tend to choose courses taught by reputedly lenient instructors, then there can be an erosion of the difficulty level of courses as students oversubscribe high-grading, easy courses relative to lower-graded, more difficult courses. This would be an educational analog of Gresham's Law in economics (counterfeit currency drives genuine currency out of circulation). Further, students will likely respond to strict instructors with low ratings, which can put pressure on those instructors to shift toward greater leniency.

Instructors who succumb to a temptation to increase grades in order to increase their ratings can be faulted for contributing to grade inflation. Although grade inflation creates problems in interpreting grades as measures of achievement, grade inflation by itself may not threaten anything fundamental to higher education. However, if increased grades are brought about by decreasing workloads — that is, by making it easier to earn high grades — then grade inflation may bring with it reduced levels of content coverage in courses. In an era in which grade inflation has been widely documented, the negative relationship between course grades and course workloads bears closer scrutiny than it has so far received.

Summary

To summarize the implications of the University of Washington research for theoretical understanding of the grades-ratings correlation: (a) the teaching effectiveness theory is more than mildly embarrassed because it does not account for anything beyond the between-class grades-ratings correlation; (b) the two third-variable theories that appeal to student motivation do better than the teaching effectiveness interpretation, because they can explain the within-class component

---

[5] A small survey of instructors in Engineering courses at University of Washington gave some support to this interpretation. Instructors who reported that they perceived their courses to be more work-demanding than typical of Engineering courses also reported that they perceived their grading policies as likely to produce a lower grade distribution than was typical of Engineering courses. A similar result from surveying faculty was reported by Marsh (1984, p. 738).

BEST COPY AVAILABLE

of the grades-ratings correlation; (c) however, two of the three third-variable theories have difficulties with the finding that grades-ratings effects are strongest when expected grades are measured in relative-grade form; (d) further, all three of the third-variable theories have difficulty both with the observed spread of the positive grades-ratings correlation to items that seem peripheral to quality of instruction, and with the observed negative relationship, across courses, between grades and workload; (e) clearly, all three of the third-variable theories are noticeably incomplete — they do not account well enough for student ratings data to provide any assurance that ratings are valid enough to be used without adjustment; (f) by contrast, the two social influence theories that credit grades with causally influencing ratings earn substantial credit by having no contradictions with the observed data; (g) further, the two grades-influence-ratings theories deserve additional credit for their ability to account for the older experimental findings that are outside the scope of all of the third-variable theories — the repeated finding that student ratings are influenced by grades manipulated in actual classroom settings.

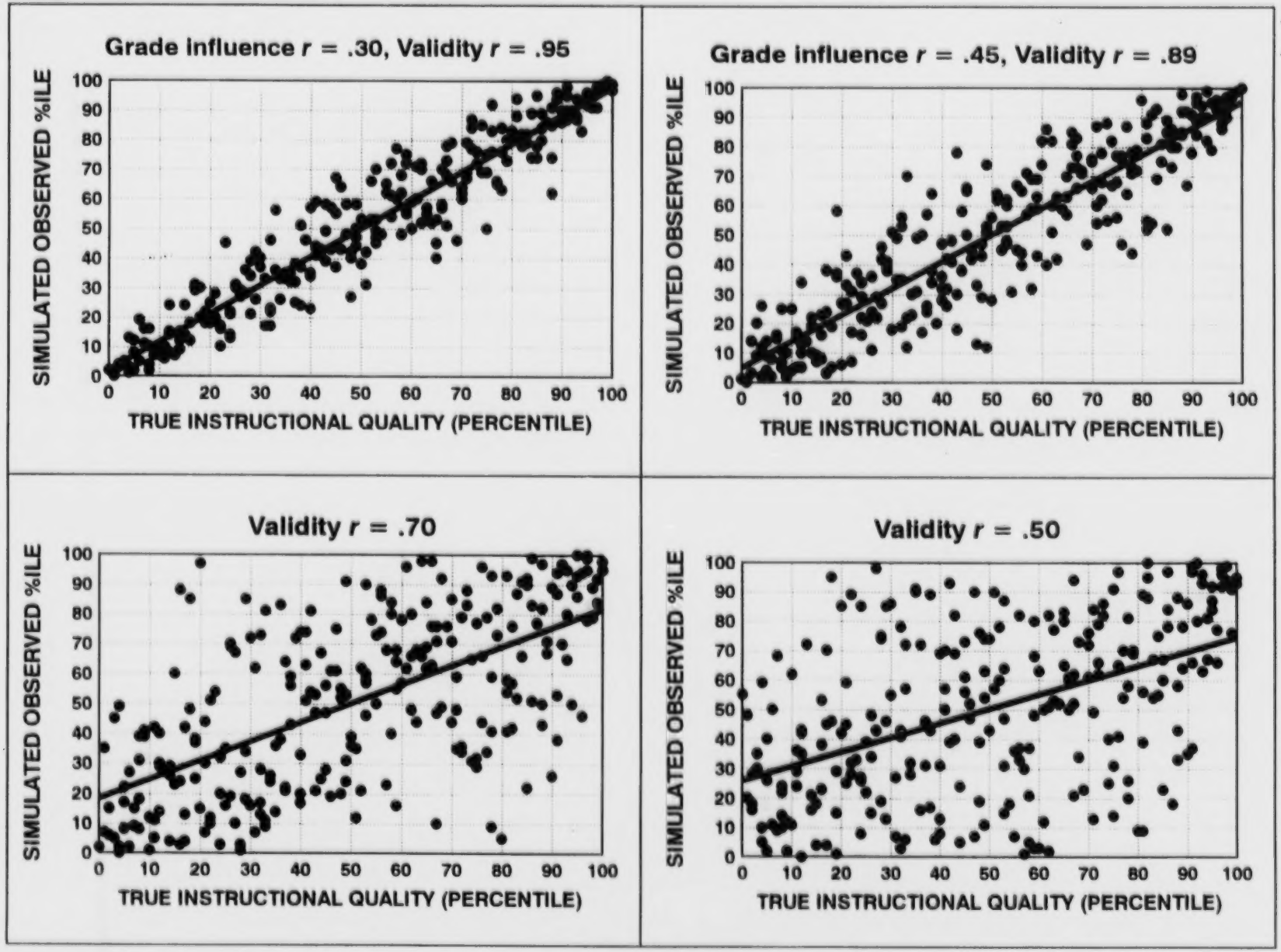
Conclusions

The findings presented here, considered in the context of much previous research on student ratings, justify the following conclusions:

1. **Giving inflated grades does produce inflated ratings**. The conclusion that grades influence ratings appears to be decisively established on the combined basis of (a) experimental studies that show impact of grades on ratings, (b) replicable correlational data patterns that are unexplained by theories that avoid the grades-influence-ratings assumption, and (c) the existence of well-established social psychological theories of interpersonal perception and judgment that predict and explain the influence of grades on ratings. The answer to the question asked earlier, *If I give higher grades, will I get higher ratings?* should be taken as a confident *yes*. The evidence certainly does not warrant the conclusion that giving high grades is, *by itself*, sufficient to assure high ratings. Nevertheless, it does support the conclusion that, if an instructor teaches varies nothing between two course offerings other than grading policy, higher ratings should be obtained in the more leniently graded section.

2. **With adjustment, student ratings may be very useful**. Their failing of discriminant validity notwithstanding, student ratings have repeatedly been shown to have modest convergent validity. In other words, at the same time that student ratings provide a distorted measure of instructional quality, they also appear to have some moderate level of valid correlation with instructional quality. The valid component of ratings may be enhanced to the extent that it is possible to statistically adjust for invalid components. This possibility is illustrated in Figures 6 and 7. Figure 6 shows four panels of simulated data. In the first two, grades are assumed to distort ratings (away from being valid measures of instructional quality) by amounts corresponding to 9% and 20% of ratings variance. These two levels of contamination correspond to grades-ratings correlations of .30 and .45, respectively. These first two panels of Figure 6 show that, even with the smaller level of contamination, there are some very substantial distortions of individual cases. In the first panel of Figure 6, for example, courses that are virtually adjacent in simulated true instructional quality are separated by as much as 40 percentile points in their simulated observed

BEST COPY AVAILABLE


Grade influence r = .30, Validity r = .95
SIMULATED OBSERVED %ILE
TRUE INSTRUCTIONAL QUALITY (PERCENTILE)
Grade influence r = .45, Validity r = .89
SIMULATED OBSERVED %ILE
TRUE INSTRUCTIONAL QUALITY (PERCENTILE)
Validity r = .70
SIMULATED OBSERVED %ILE
TRUE INSTRUCTIONAL QUALITY (PERCENTILE)
Validity r = .50
SIMULATED OBSERVED %ILE
TRUE INSTRUCTIONAL QUALITY (PERCENTILE)

BEST COPY AVAILABLE

ratings. The first two panels of Figure 6 assume very high validity coefficients for the ratings measure (.95 and .89, respectively). However, validity correlations with nonratings measures of instructional quality are rarely found to be above .50, rather averaging about .40 (in the multisection validity designs). The lower two panels of Figure 6 simulate the relationship of ratings to true instructional quality when more realistic (but still high) validity coefficients of .70 and .50 are assumed. With the level of noise in the system modeled by Figure 6's simulations, any possibility for eliminating systematic portions of the invalid variance should be pursued.

Figure 6. Simulated discrepancy between a hypothetical construct of true instructional quality and observed ratings for four levels of assumed validity of ratings. The first two panels show the effect of assuming that the *only* source of invalidity is contamination by grades-ratings influences that explain 9% and 20% of variance in ratings, respectively. The first two panels assume implausibly high validity correlations of .95 and .89, respectively. The next two panels assume validity correlations that, although lower than those in the first two panels, are still higher than those demonstrated for acutal ratings data. With the levels of validity shown in the last two panels, courses that are very similar in true instructional quality can have extremely divergent observed ratings.

Figure 7 illustrates a set of actual ratings data to which adjustments for grades-ratings correlation and a few lesser influences have been applied. These adjustments can be seen to have shifted the relative standing of courses up or down by more than three deciles for about 10% of the sample of courses. Note, for example, that courses very near the median before adjustment are distributed from the highest to the lowest decile after adjustment.

Figure 7. Example of adjustment of actual ratings data to reduce likely contamination associated with the following measures (percentages of rating variance associated with each shown in parentheses): grading policies (combined relative and absolute grade: 27.3%), class level (freshman to graduate) and enrollment (combined: 3.6%). Data from University of Washington, Winter Quarter, 1994 (N = 254 courses; same sample as in Figure 2).

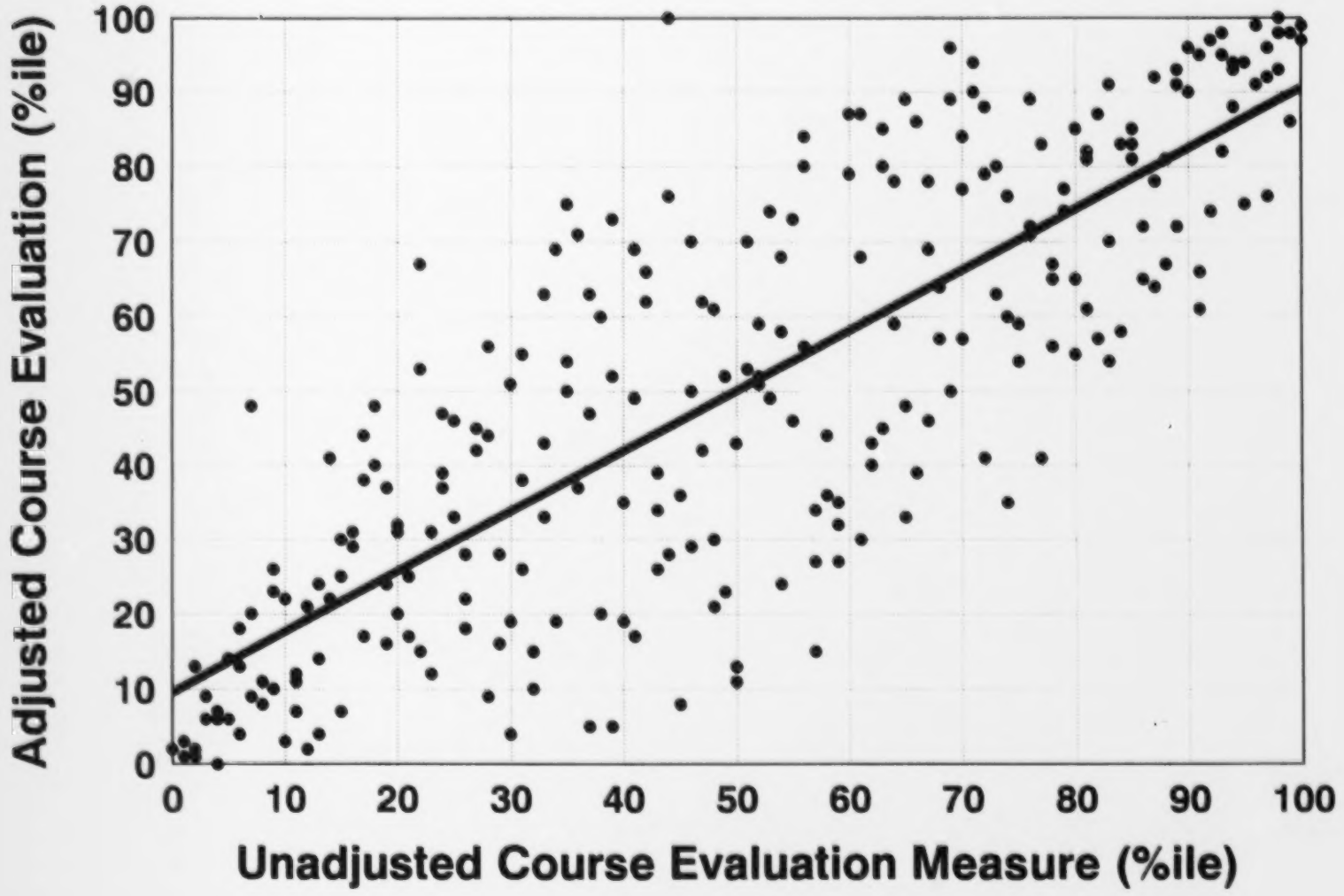
3. **Workload measures are useful**. The consistent finding of a negative relationship between course grades and workload (illustrated in Figure 5) is disturbing. Although this relationship may exist at many colleges and universities, it has never become a focus of research attention, perhaps because workload measures are not included in many course rating forms. The inclusion of workload estimates in course evaluation forms can assure that this important aspect of differences among courses does not continue to escape attention.

The Baby and the Bathwater

This examination of psychological processes underlying student ratings might be interpreted as sufficient basis for abandoning the whole enterprise of conducting student ratings. However, there are three good reasons to conclude just the reverse — that even more attention should be paid to ratings.

First, in many cases there is no practical alternative method for evaluating instruction. Although expert appraisals and standardized achievement tests might, in principle, provide more valid

BEST COPY AVAILABLE


Adjusted Course Evaluation (%ile)
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
Unadjusted Course Evaluation Measure (%ile)

assessments, unfortunately both of those alternative methods are considerably more costly than student ratings. Their present very limited use probably stands as an appropriate indicator of their relative impracticality.

Second, the evidence for convergent validity of student ratings should not be dismissed. Although student ratings are overlaid with some misleading artifacts, they nevertheless also contain useful information. Theory-based statistical adjustments can make that information more usable than it presently is.

Third, even a worst-case scenario suggests that student ratings can provide useful information. In this worst-case scenario, one might conclude that adjusted student ratings provide information only about how well students like a course, and nothing at all about how much students are learning from the course. Still, this assessment of liking or attitude should be very useful, in the same way that an assessment of bedside manner is useful in evaluating a physician. The assessment of bedside manner doesn't describe the physician's success in preventing or curing illness, but it does give information that may predict a patient's willingness to adhere to prescribed treatments and to return for future checkups. Similarly, knowledge of how much a teacher is liked should provide information that can predict a student's willingness to do assigned work and to register for further course work from that teacher.

In summary, there very likely *is* an instructional quality baby in with the bathwater of grades-ratings correlations and other possible contaminants of ratings. It seems much, much wiser to give that baby a bath, to clean it up and make it presentable, than to abandon the baby in the process of discarding the bathwater.

BEST COPY AVAILABLE

REFERENCES

Abrami, P. C., Cohen, P. A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research, 58*, 151-179.

Abrami, P. C., Dickens, W. J., Perry, R. P., & Leventhal, L. (1980). Do teacher standards for assigning grades affect student evaluations of instruction? *Journal of Educational Psychology, 72*, 107-118.

Aronson, E., & Linder, D. E. (1965). Gain and loss of esteem as determinants of interpersonal attractiveness. *Journal of Experimental Social Psychology, 1*, 156-171.

Cashin, W. E. (1995). Student ratings of teaching: The research revisited. IDEA Paper No. 32. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Chacko, T. I. (1983). Student ratings of instruction: A function of grading standards. *Educational Research Quarterly, 8*(2), 19-25.

Feldman, K. A. (in press). Identifying exemplary teachers and teaching: Evidence form student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice*. New York: Agathon Press.

Freedman, R. D., Stumpf, S. A., & Aguanno, J. C. (1979). Validity of the Course-Faculty Instrument (CFI): Intrinsic and extrinsic variables. *Educational & Psychological Measurement, 39*, 153-158.

Gigliotti, R. J., & Buchtel, F. S. (1990). Attributional bias and course evaluations. *Journal of Educational Psychology, 82*, 341-351.

Gillmore, G. M., & Greenwald, A. G. (1994, April). The effects of course demands and grading leniency on student ratings of instruction. Paper presented at meetings of American Educational Research Association, Orlando, FL.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American Psychologist, 35*, 603-618.

Greenwald, A. G., & Gillmore, G. M. (in preparation). Use of measures of expected grade to assess validity of student ratings of instruction. University of Washington.

Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *Journal of Educational Psychology, 63*, 130-133.

Howard, G. S., Conway, C. G, & Maxwell, S. E. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology, 77*, 187-196.

Howard, G. S., & Maxwell, S. E. (1980). Correlation between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology*, *72*, 810-820.

Howard, G. S., & Maxwell, S. E. (1982). Do grades contaminate student evaluations of instruction? *Research in Higher Education*, *16*, 175-188.

Knowles, E. S. (1982). A comment on the study of classroom ecology: A lament for the good old days. *Personality and Social Psychology Bulletin*, *8*, 357-361.

Marsh, H. W. (1980). The influence of student, course, and instructor characteristics on evaluations of university teaching. *American Educational Research Journal*, *17*, 219-237.

Marsh, H. W. (1982). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology*, *74*, 264-279.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76*, 707-754.

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. *Higher Education: Handbook of Theory and Research*, *8*, 143-233.

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe*, *65*, 384-397.

Powell, R. W. (1977). Grades, learning, and student evaluation of instruction. *Research in Higher Education*, *7*, 193-205.

Snyder, C. R., & Clair, M. (1976). Effects of expected and obtained grades on teacher evaluation and attribution of performance. *Journal of Educational Psychology*, *68*, 75-82.

Stumpf, S. A., & Freedman, R. D. (1979). Expected grade covariation with student ratings of instruction: Individual versus class effects. *Journal of Educational Psychology*, *71*, 293-302.

Theall, M., Franklin, J., & Ludlow, L. (1990). Attributions and retributions: Student ratings and the perceived causes of performance. *Instructional Evaluation*, *11*, 12-17.

Vasta, R., & Sarmiento, R. F. (1979). Liberal grading improves evaluations but not performance. *Journal of Educational Psychology*, *71*, 207-211.

Worthington, A. G., & Wong, P. T. P. (1979). Effects of earned and assigned grades on student evaluations of an instructor. *Journal of Educational Psychology*, *71*, 764-775.

Appendix

INSTRUCTIONAL ASSESSMENT SYSTEM

Office of Educational Assessment
University of Washington

Fill in bubbles darkly and completely
Erase errors cleanly

FORM X

Instructor _____ Course _____ Section _____ Date _____

Completion of this questionnaire is voluntary. You are free to leave some or all questions unanswered.

How frequently was each of the following a true description of this course?	Always			About Half			Never
1. The instructor gave very clear explanations.	○	○	○	○	○	○	○
2. The instructor successfully rephrased explanations to clear up confusion.	○	○	○	○	○	○	○
3. Class sessions were interesting and engaging.	○	○	○	○	○	○	○
4. Class sessions were well organized.	○	○	○	○	○	○	○
5. Student participation was encouraged.	○	○	○	○	○	○	○
6. Students were aware of what was expected of them.	○	○	○	○	○	○	○
7. Extra help was readily available.	○	○	○	○	○	○	○
8. Assigned readings and other out-of-class work were valuable.	○	○	○	○	○	○	○
9. Grades were assigned fairly.	○	○	○	○	○	○	○
10. Meaningful feedback on tests and other work was provided.	○	○	○	○	○	○	○
11. Evaluation of student performance was related to important course goals.	○	○	○	○	○	○	○

Relative to other college courses you have taken, how would you describe your progress in this course with regard to:	Great			Average			None
12. Learning the conceptual and factual knowledge of this course.	○	○	○	○	○	○	○
13. Developing an appreciation for the field in which this course resides.	○	○	○	○	○	○	○
14. Understanding written material in this field.	○	○	○	○	○	○	○
15. Developing an ability to express yourself in writing or orally in this field.	○	○	○	○	○	○	○
16. Understanding and solving problems in this field.	○	○	○	○	○	○	○
17. Applying the course material to real world issues or to other disciplines.	○	○	○	○	○	○	○
18. General intellectual development.	○	○	○	○	○	○	○

Relative to other college courses you have taken:	Much Greater			Average			Much Less
19. Do you expect your grade in this course to be:	○	○	○	○	○	○	○
20. The intellectual challenge this course presented was:	○	○	○	○	○	○	○
21. The amount of effort to succeed in this course was:	○	○	○	○	○	○	○
22. Your involvement in this class (doing assignments, attending classes, etc.) was:	○	○	○	○	○	○	○

If you had it to do over again and this course was optional for your program, would you enroll in it:	Certainly			Neutral			Certainly Not
23. If the same instructor taught it?	○	○	○	○	○	○	○
24. If a different instructor taught it?	○	○	○	○	○	○	○
25. Regardless of who taught it?	○	○	○	○	○	○	○

26. On average, how many hours per week have you spent on this class, including attending classes, doing readings, reviewing notes, writing papers and any other course related work?

○ Under 2	○ 6-7	○ 12-13	○ 18-19
○ 2-3	○ 8-9	○ 14-15	○ 20-21
○ 4-5	○ 10-11	○ 16-17	○ 22 or more

27. From the total average hours above, how many do you consider were valuable in advancing your education?

○ Under 2	○ 6-7	○ 12-13	○ 18-19
○ 2-3	○ 8-9	○ 14-15	○ 20-21
○ 4-5	○ 10-11	○ 16-17	○ 22 or more

28. What grade do you expect in this class?

○ A (3.8-4.0)	○ B (2.8-3.2)	○ C (1.8-2.2)	○ D (0.7-1.2)	○ Credit
○ A- (3.6-3.7)	○ B- (2.6-2.7)	○ C- (1.6-1.7)	○ E (0.0)	○ No Credit
○ B+ (3.3-3.5)	○ C+ (2.3-2.5)	○ D+ (1.3-1.5)	○ PASS	

29. In regard to your academic program is this course best described as:

○ In your major?	○ A distribution requirement?	○ An elective?	
○ In your minor?	○ A program requirement?	○ Other?	

30. What is your current class standing?

○ Freshman	○ Junior	○ Graduate	○ Other
○ Sophomore	○ Senior	○ Professional	

©CEA 1993

BEST COPY AVAILABLE